基于主题模型的百科知识库实体对齐 *

刘振鹏 [a, b]，贺梦洁 [a]，张　彬 [b]，董　静 [a]，徐建民 [c]

(河北大学 a. 电子信息工程学院; b. 信息技术中心; c. 网络空间安全与计算机学院, 河北 保定 071002)

摘　要：针对传统实体对齐的方法无法体现潜在语义信息的问题，对其进行优化，使实体对齐效果更加显著。使用 LDA 模型对网络百科非结构化数据进行建模，采用改进的 BP 算法求解 LDA 模型中的隐藏参数，进而生成实体特征向量进行相似度计算，通过计算结果判断是否可以对齐。实验结果表明，通过与三种传统的算法进行比较，所提算法在准确率、召回率和综合指标 F 值三个评价指标均有所提高。针对具有描述信息的网络百科实体，该算法可以有效提升实体对齐效果。

关键词：实体对齐；LDA 模型；BP 算法；知识融合

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.05.0305

Entity alignment for encyclopedia knowledge base based on topic model

Liu Zhenpeng[a, b], He Mengjie[a], Zhang Bin[b], Dong Jing[a], Xu Jianmin[c]

(*a. School of Electronic Information Engineering*, *b. Information Technology Center*, *c. School of Cyber Security & Computer Hebei University*, *Baoding Hebei* 071002, *China*)

Abstract: Aiming at the problem that traditional entity alignment method could not reflect latent semantic information, it was optimized, making the effect of entity alignment more significant. Using the LDA model to model the unstructured data of the network encyclopedia, and with the improved BP algorithm to solve the hidden parameters of LDA model, in turn, generate entity eigenvectors to perform similarity calculation, finally, through calculation results can determine whether alignment. The results showed that, through comparing with three kinds of traditional algorithms, the algorithm which proposed in this paper have increased the three evaluation index that above Precision, Recall and F-score. Aiming at the network encyclopedia entity with description information, the algorithm can effectively improve the entity alignment effect.

Key words: entity alignment; LDA model; BP algorithm; knowledge fusion

0　引言

近十几年，互联网产生了越来越多的大规模知识库，例如国外具有代表性的知识库 FreeBase[1]，DBpedia[2]，维基百科本体知识库（yet another great ontology,YAGO[3]）和 Omega[4]等；在我国，著名知识库有百度知心，搜狗知立方及清华大学双语知识库 XLore[5]。知识库在知识图谱[6]，信息融合及智能语义问答[7]等自然语言处理和人工智能领域均有重要的意义。中文知识库构建中，可用的完备数据资源比较少，在获取完整的知识的过程中，需要将不同知识库里的知识数据进行集成、整合和复用，实体对齐作为知识融合的重要方法对知识库的构建和扩充产生着重要的作用。

实体[8]（entity）是指客观存在并且可以进行区别的事物，包括具体的人，事，物，抽象的概念或关系等。实体对齐（entity alignment），也可被称为是实体链接[9]，其目的是判断不同数据源[10]中的两个实体是否指向现实世界中的同一对象。

目前，实体对齐方法的研究主要有基于网络本体语义[11]（Web ontology language，OWL），基于规则分析，基于相似度理论判定三种。针对中文网络百科，它本身不具有完备的本体信息，因此，它很难通过 OWL 语义进行对齐；并且网络百科当中包含的实体领域众多，若通过建立规则进行对齐，不同的领域要建立不同的规则，这类方法又不具有通用性；使用比较广泛的是基于相似度理论进行判定，通常，这一类的方法通过对属性值赋予权重[12]，然后通过计算不同实体的同一属性的相似度进行实体对齐，近几年由于主题模型的盛行也出现了应用主题模型对实体的描述性文本进行建模，之后运用相似度进行实体对齐的方法。文献[13,14]利用 RDFS 词表对属性进行规范化之后，利用属性相似度和描述性文本的主题特征相似度进行

收稿日期：2018-05-24；**修回日期：**2018-07-09　　**基金项目：**河北省自然科学基金资助项目（2015201142）

作者简介：刘振鹏（1966-），男，河北安国人，教授，博士，主要研究方向为大数据、网络信息安全、自然语言处理；贺梦洁（1992-），女，硕士研究生，主要研究方向为大数据、自然语言处理；张彬（1980-），男（通信作者），高级工程师，硕士研究生，主要研究方向为网络安全（zb@hbu.edu.cn）；董静（1992-），女，硕士研究生，主要研究方向为大数据、自然语言处理；徐建民（1966-），男，教授，博导，主要研究方向为信息检索、不确定信息处理.

结合，实现了实体对齐；文献[15]提出一种半监督协同训练的实体对齐方法，结合实体名称、属性、描述文本及其中的时间、数值等关键的信息进行实体对齐；文献[16]提出一种独立于本体模式的基于属性语义特征的实体对齐方法，采用的仍然是实体的属性信息。然而这样的方法对于匮乏属性信息的实体则不适用，尤其对于中文网络百科，不同网络百科的相同属性的名称甚至属性信息出现了很多不一致的情况，例如众所周知的百度百科和互动百科这两个国内规模较大的网络百科网站，在“英文名”这一属性项目中，百度百科采用的是“外文名”，而互动百科采用的则是“英文名”；而对于歌手“张杰”这一公众人物的“别名”这一属性项，百度百科采用的的是“杰哥”，而互动百科采用的是“张小杰”，这种现象对于采用属性信息进行实体对齐无疑是增加了一定的难度，在这个过程中首先要考虑的就是对于属性的名称进行统一，若无法保证属性对齐的准确率，则对于最后的结果有很大的影响，并且通过研究，对于中文网络百科而言，属性信息在处理不当的情况下会产生不良效果，并且加大了实体对齐的工作量。因此，百科知识库中包含的大量实体摘要信息和描述性文本可以被有效利用，如何只利用实体的非结构化文本构造出可以有效的进行实体对齐的实体特征是本文所面对的问题。

为了有效的利用实体非结构化文本，本文提出了基于主题模型的百科知识库实体对齐算法，该算法利用 LDA 模型对网络百科实体的文本信息进行主题建模，使用改进的 BP 算法求解模型中的隐藏参数，进而完成实体对齐任务。经过实验证明，所提方法能够有效的提高实体对齐的准确率，对具有描述性文本的实体进行实体对齐有很好的通用性。

本文主要工作如下：a）有效的利用百科实体的非结构化数据，使用 LDA 模型得到文本中潜在的语义信息，提出一种广泛适用于具备描述信息的百科实体对齐算法;b）在推断 LDA 模型隐藏参数的时候，提出改进的 BP 算法对模型参数进行估计;c）获取百度百科和中文维基百科数据进行实验验证，与同类相似算法进行对比，并对算法的有效性进行分析。

1 相关知识介绍

1.1 LDA 模型

潜在狄利克雷分配[17]（Latent Dirichlet Allocation，LDA）是由 Blei 等人在 2003 年提出的一种三层贝叶斯概率模型，它包括单词、主题、文档三层。图 1 是 LDA 图模型。

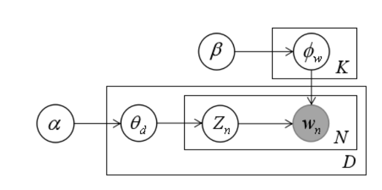


图 1 LDA 图模型

图 1 中，白色圆圈表示隐藏变量，灰色圆圈表示可以观测到的变量，圆圈之间的箭头表示两个变量之间的概率是相关的，方框代表重复，方框里的下标是重复的次数。α 和 β 分别表示两个分布 θ_d 和 ϕ_k 的先验参数，在本文及实验当中 α 和 β 取值均为 0.1；w_n 表示文档中的某个单词，Z_n 表示文档中某个单词 w_n 的主题，K 表示主题的总个数，N 表示文档中词的个数，D 表示文档篇数。图 1 中 LDA 图模型是从文档生成的角度对该模型进行介绍，也就是单词 w_n 被选择出来的过程。

该模型将文本生成的步骤简化为概率采样的步骤，将文档表示为多个主题的概率混合即“文档—主题”概率矩阵 θ_d，而主题又可以由不同的单词构成，即“主题—单词”概率矩阵 ϕ_k，因此要生成一篇文章，先是对主题进行采样，从而得到了该主题下的单词集合，进行迭代抽取多个单词，从而得到完整的文章。

针对本文所涉及的算法，需要对模型中出现的两个参数 θ_d 和 ϕ_k 进行参数估计，从而进行实体对齐实验。目前主流的参数估计方法有三种，变分贝叶斯（variational Bayesian，VB），吉布斯采样（Gibbs Sampling，GS）和置信传播（Belief Propagation，BP），虽然变分贝叶斯算法和吉布斯采样在近似推理方面取得了不小的进展，基于 BP 算法在学习速度和准确率的方面均有很强的竞争力，本文所提算法中采用经典的神经网络置信传播（belief propagation，BP）算法并对其进行优化。

1.2 置信传播算法

BP 算法是由 Pearl[18]提出的一种推断图模型参数的信息传递算法，是一种有效求解条件件边缘概率的方法，Zeng 等人[19]在 2011 年将该算法应用到求解 LDA 模型隐藏变量，即求解 θ_d 和 ϕ_w 的值，并取得了很大的进展。图 2 是由 Zeng Jia 提出的基于 BP 算法的 LDA 因子图。

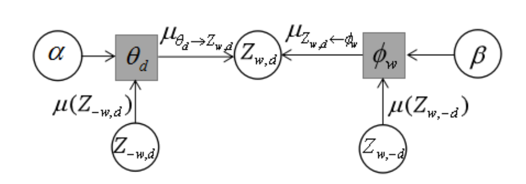


图 2 基于 BP 算法的 LDA 因子图

它与图 1 的 LDA 图模型是一个模型，只是图 1 侧重于文档生成部分，图 2 则是侧重于各个主题标签之间的关系，并且凸显了主题标签求解的数学关系。在图 2 中，灰色的方框表示需要求解的隐藏变量 θ_d 和 ϕ_w，α 和 β 仍然是表示 θ_d 和 ϕ_w 的先验参数，其他的则是表示主题标签；θ_d 连接 $Z_{w,d}$ 和 $Z_{-w,d}$，$Z_{w,d}$ 表示文本 d 中单词 w 的主题标签，$Z_{-w,d}$ 表示文本 d 中除了单词 w 之外的其他单词的主题标签，也就是说 θ_d 连接着同一文本 d 中的所有单词的主题标签，而 ϕ_w 连接的是 $Z_{w,d}$ 和 $Z_{w,-d}$，$Z_{w,d}$ 如上，$Z_{w,-d}$ 指除了当前文本 d 之外的所有其他文本中单词 w 的主题标签，那么 ϕ_w 连接的就是全部文本中单词 w 的主题标签。图 2 中箭头表示的是信息的传递方向，箭头上承载的便是主题所包含的信息，即 $\mu_{\theta_d \to Z_{w,d}}$，$\mu_{Z_{w,d} \leftarrow \phi_w}$，$\mu(Z_{-w,d})$ 和 $\mu(Z_{w,-d})$ 所表示的是该模型的主题信息。

使用 BP 算法对 LDA 进行参数估计时，对 $Z_{w,d}$ 有影响的是与其相连的所有的主题标签与参数，其主题更新公式为

$$\mu(Z_{w,d}=k)\propto\frac{\mu(Z_{-w,d}=k)+\alpha}{\sum_{k}[\mu(Z_{-w,d}=k)+\alpha]}\times\frac{\mu(Z_{w,-d}=k)+\beta}{\sum_{w}[\mu(Z_{w,-d}=k)+\beta]} \tag{1}$$

其中：$\mu(Z_{-w,d}=k)$ 表示的是文档 d 中除了单词 w 以外的其余所有单词的主题概率分布，而 $\mu(Z_{w,-d}=k)$ 表示的是除了文档 d 其余的所有文档中单词 w 的主题概率分布；

$$\mu(Z_{-w,d}=k)=\sum_{-w}x_{-w,d}\mu(Z_{-w,d}=k) \tag{2}$$

$$\mu(Z_{w,-d}=k)=\sum_{-d}x_{w,-d}\mu(Z_{w,-d}=k) \tag{3}$$

其中：$x_{w,d}$ 表示的是观测值。在式（2）中 $x_{-w,d}$ 表示文档 d 中除了单词 w 以外其余全部单词的观测值，式（3）中 $x_{w,-d}$ 表示的是除了文档 d 其余的所有文档中单词 w 的观测值。

其中，信息更新被局部归一化，即 $\sum_{k}\mu(Z_{w,d}=k)=1$，其待估计参数 θ_d 和 ϕ_w：

$$\theta_d(k)=\frac{\mu(Z_{\cdot,d}=k)+\alpha}{\sum_{k}[\mu(Z_{\cdot,d}=k)+\alpha]} \tag{4}$$

$$\phi_w(k)=\frac{\mu(Z_{w,\cdot}=k)+\beta}{\sum_{w}[\mu(Z_{w,\cdot}=k)+\beta]} \tag{5}$$

其中：$\mu(Z_{\cdot,d}=k)$ 表示文档 d 中所有单词的主题概率分布；$\mu(Z_{w,\cdot}=k)$ 表示所有文档中单词 w 的主题概率分布。

2　算法实现

2.1　算法概述

本文核心任务是针对具有相同条目名称的百科实体，计算它们之间的潜在语义的相似度，对实体进行对齐。具体算法过程如图 3 所示。

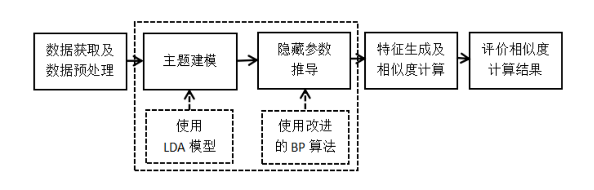


图 3　基于主题模型的百科知识库实体对齐算法

由图 3 可知，该算法包含四个模块，第一个模块是数据获取和数据预处理，在这一部分，本文获取了维基百科中文版的语料和部分百度百科的语料，这些语料当中包括百科实体的条目名称和相关的描述信息，数据获取之后，对其进行分词和去停用词处理；之后对处理好的文本使用 LDA 模型进行主题建模，然后，使用改进的 BP 算法对得到的 LDA 模型进行参数估计，这一部分是该算法的核心步骤，将在 2.2 进行详细的介绍；再次是特征生成和相似度计算模块，特征生成过程是对得到的“文档—主题”矩阵 θ_d 进行处理以得到实体的特征向量，而相似度计算则采用余弦相似度进行计算，这一部分将在 2.3 进行介绍；最后是评价相似度计算结果，计算两个实体之间的相似度大于阈值 ω 时，则判定为可对齐，否则，待对齐实体 *NE* 则作为新的实体保存在实体库中，并添加到候选实体的义项当中。

2.2　改进的 BP 算法

传统的 LDA 模型是基于词袋模型，于是单词之间的顺序就不被考虑，这样的做法使模型变得简单，但是也为其改进提供了机会[20]。BP 算法在对 LDA 模型进行参数推断时有精度高、速度快的优势，但是由于 LDA 模型本身的缺陷，以及该算法主要是针对中文网络百科，而中文当中的单词词义大概率是要根据上下文理解的，由此，提出一种改进的 BP 算法。图 4 是改进的 BP 算法因子图。

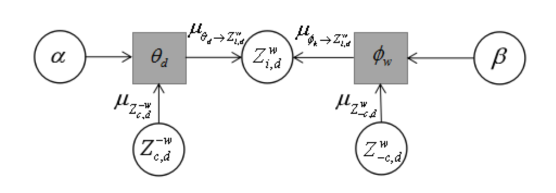


图 4　改进的 BP 算法因子图

在改进的 BP 算法当中，新添加了一项内容，即单词 w 的上下文（用字母 c 来表示），也就是在计算单词的主题分布的时候，将这个单词作为中心，在它的前后扩展若干个单词形成一个单词集窗口，以这个窗口为短文本来计算每个单词的主题分布，最后经过迭代，使每个单词的主题分布达到收敛。其中，$Z_{i,d}^w$ 表示在文档 d 中第 i 个单词 w 的主题标签；$Z_{c,d}^{-w}$ 表示在文档 d 中，除了单词 w 外的上下文其他单词的主题标签，$Z_{-c,d}^w$ 表示文档 d 中，除了上下文外其他的单词 w 的主题标签；此外 α，β，θ_d 和 ϕ_w 表示的内容与 BP 算法一致。

该算法对于 BP 算法的优化首先是将上下文的概念引入本算法，这一点主要是针对 LDA 模型当中单词之间的顺序不被考虑并且 BP 算法为同一文档中的相同单词分布了相同的语义信息的缺陷，加入上下文之后，同一篇文档中的单词顺序则不会被打乱，并且针对中文的特点理解一个单词要结合其上下文去理解，则同一篇文档中相同的单词则会为其分布不同的主题，使单词的语义更加贴近其真实语境。其次，本次改进还将 $Z_{w,-d}$ 这一项改为 $Z_{-c,d}^w$ 则是因为本文的目的是通过比较两个具有相同名称的百科实体的描述信息来进行实体对齐，若将其他的文档中相同单词的主题信息加入，则会对本篇文档的主题信息造成混淆，因此，本文采用了 $Z_{-c,d}^w$，等同于只采用同一篇文档中的相同单词的信息，会使文档主题更加明确。

由该算法的因子图可以看出，文档 d 中第 i 个单词 w 的主题由以下两个部分决定，一是上下文窗口中不同单词的主题影响，二是同一文档中非上下文窗口中相同单词主题的影响。由此得到主题更新公式为

$$\mu(Z_{i,d}^w=k)\propto\frac{\mu(Z_{c,d}^{-w}=k)+\alpha}{\sum_{k}[\mu(Z_{c,d}^{-w}=k)+\alpha]}\times\frac{\mu(Z_{-c,d}^w=k)+\beta}{\sum_{w}[\mu(Z_{-c,d}^w=k)+\beta]} \tag{6}$$

该公式中 $\mu(Z_{c,d}^{-w}=k)$ 表示的是上下文中除了单词 w 外其他单词的主题信息；$\mu(Z_{-c,d}^w=k)$ 表示同一文档 d 中除了上下文以外，单词 w 的主题信息。

最后得到模型的参数为

$$\theta_d \propto \frac{\mu(Z_{c,d}^{-w}=k)+\alpha}{\sum_k[\mu(Z_{c,d}^{-w}=k)+\alpha]} \tag{7}$$

$$\phi_w \propto \frac{\mu(Z_{-c,d}^{w}=k)+\beta}{\sum_w[\mu(Z_{-c,d}^{w}=k)]+\beta} \tag{8}$$

根据以上的因子图和主题更新公式，使用改进的 BP 算法估计 LDA 模型的隐藏参数的训练过程为：

a)随机为每个单词初始化一个主题；

b)遍历整个语料库，使主题更新式（6）更新每个单词的主题分布；

c)不断迭代上述过程直至收敛；

d)使用式（7）（8）求出参数。

2.3 特征向量生成和相似度计算

在使用 LDA 进行主题建模时，文本的主题是隐藏变量，也就是 θ_d 和 ϕ_w 的值是未知的，本文使用改进的 BP 算法对模型的未知参数进行估计。

1) 计算“文档—主题”概率矩阵 θ_d

通过对实体的描述信息进行主题建模，使用改进的 BP 算法对模型的隐藏变量进行估计得到“文档—主题”概率矩阵 θ_d：

$$\theta_d = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nK} \end{bmatrix} \tag{9}$$

其中：p_{ij} 表示第 j 个主题归入第 i 个文档的概率；n 表示文档集中有 n 篇文档，K 表示 LDA 进行建模时生成了 K 个主题。

2）将 θ_d 按行进行拆分生成“文档—主题”向量

输入的文档集合 $D=(d_0,d_1,\cdots,d_n)$ 其中 d_0 表示待对齐文本，其余文本表示实体库中的条目名称相同的文本。

其中：$d_0=(p_{11},p_{12},\cdots,p_{1k})$，$d_1=(p_{21},p_{22},\cdots,p_{2k})$，…$d_n=(p_{n1},p_{n2},\cdots,p_{nk})$

3）相似度计算

将 d_0 与其余的“文档—主题”向量进行余弦相似度的计算以求出两篇名称相同的文章的文档相似度。例如由 d_0 代表的实体 e_a 的“文档—主题”向量和另一实体 e_b 的主题相似度为

$$sim(e_a,e_b)=\frac{d_0\cdot d_i}{|d_0||d_i|} \tag{10}$$

其中: d_i 表示实体 e_b 的“文档—主题”向量。

3 实验

3.1 实验数据

为了验证本文所提算法的有效性，本文采用质量相对较高的中文语料库维基百科中文版和百度百科的文本数据进行实验。维基百科会定时将自己的语料库进行更新并打包发布，本文下载了最新的维基语料进行实验，语料部分包括词条名称和相应的描述信息。由于维基百科语料较为全面并且包含信息很多，维基百科的语料库在本实验中作为实体库存在。而百度百科的语料则需要进行爬取，本文在百度百科网站爬取了人物类，社会类，科学类和艺术类各 200 条，共 800 条百度百科的词条信息，其中包括词条名称和相应的描述信息，作为待对齐实体进行实验。

获取到实验数据之后，进行数据预处理，本文利用 Python 语言进行实验，在数据预处理部分使用 Python 自带的 jieba 分词进行分词处理，使用“哈工大停用词表”进行去除停用词。本文实验所用数据统计如表 1 所示。

表 1 实体对齐数据统计

分类	百度百科实体数	维基百科重名实体数	可对齐数
人物	200	3897	188
社会	200	635	120
科学	200	585	126
艺术	200	1149	159

表 1 概括统计了本文实验用到的数据量，如表 1 所示，本文实验在百度百科网站从人物、社会、科学和艺术四类条目分别爬取了 200 条百科实体，并按照条目名称进行抽取，获取到在维基百科中的同名实体，并得到了其个数，与此同时，经过人工比对，还得到了可对齐数。

3.2 评价标准

实验的评价标准从准确率（precision，P），召回率（recall，R）以及综合指标 F 值（F-score，F）[21]三项来进行评价，其中，

a)准确率（P）,表示经过实体对齐算法后得到准确对齐的数量和参与对齐的实体数的比率

$$P=N_r/N_o \tag{11}$$

b)召回率（R）,表示经过算法之后准确对齐的数量和数据集当中可对齐实体的比率

$$R=N_r/N_a \tag{12}$$

c)综合指标 F 值（F）,表示权衡准确率和召回率的综合指标

$$F=2\cdot P\cdot R/(P+R) \tag{13}$$

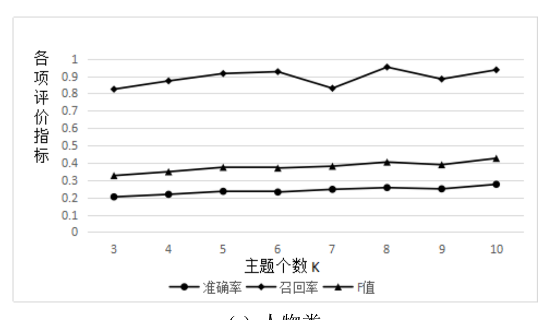
其中: N_r 表示经过本算法之后准确对齐的实体数，N_o 表示在本次实验中参与对齐的实体数，N_a 表示数据集当中可以准确对齐的实体数。

3.3 参数设定

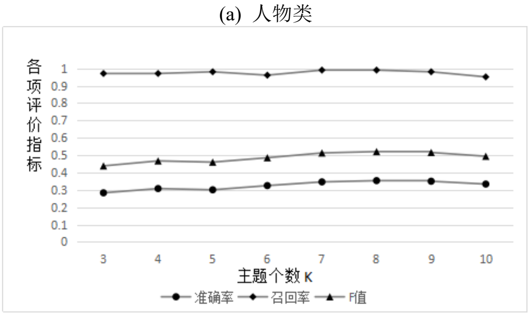
本文的参数主要有三个，首先是针对 LDA 模型，需要设定其主题个数 K；其次则是针对改进的 BP 算法，即在求解隐藏参数过程中需要设定其先验参数 α 和 β，经过对先前的研究成果的借鉴以及进行相应的实验，本文把先验参数设定为统一的 $\alpha=0.1$ 和 $\beta=0.1$，由于他们并不是本文的研究重点，因此本文便略过了对它的推理过程，有关于先验参数的设定，Wallach 等人[22]提出了较多有效理论；最后是针对本文所提基于 LDA 的实体对齐算法当中的阈值 ω 的设定，这一参数的设定与实体对齐的结果直接相关，因此会对这一部分作详细的介绍。

1)主题个数 K 对实验结果影响

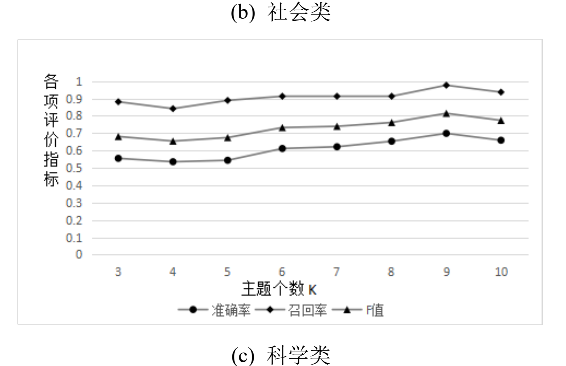
为了避免阈值对实验影响，在这一部分实验中阈值ω设定为0.9。实验结果如图5所示。



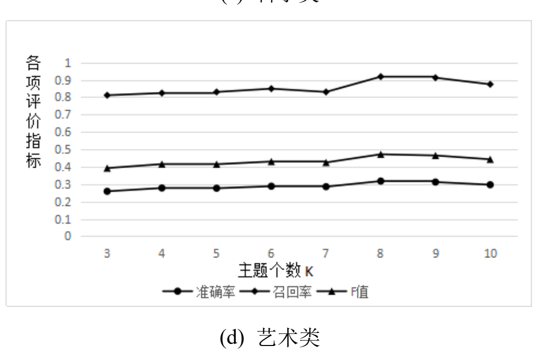
(a) 人物类



(b) 社会类



(c) 科学类



(d) 艺术类

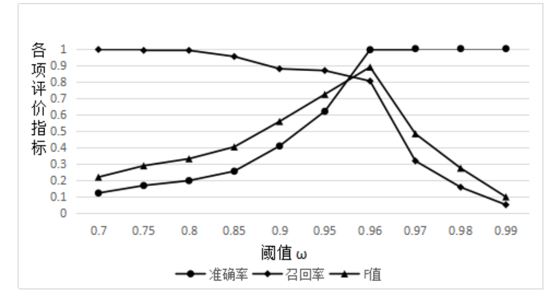
图5 主题个数K对实验结果的影响

由以上的实验结果可以看出，当主题个数K为8或者是9时实体对齐的准确率（P），召回率（R）和综合指标F值（F）均为最优。

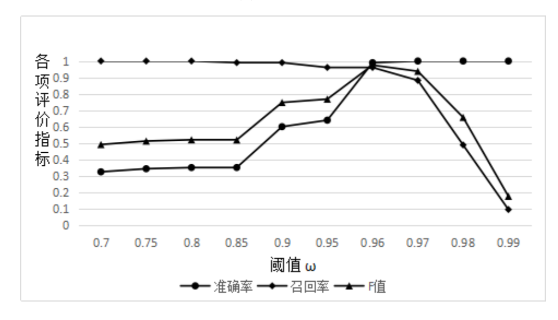
由图5可知，人物类、社会类和艺术类的实体对齐准确率均不是很高，有较大可能是因为这三类实体的描述信息不够准确，并且数据显示实体库中这三类实体的同名实体数目较大，从而增加了参与对齐的实体数目N_o，而科学类的实体由于其描述信息较为严谨、清晰，其准确率相对来说比较高，并且其专有名词较多，并不很容易出现同名实体，其参与对齐的实体数目N_o较少，因此同样的主题个数的条件下，其准确率较高。

2)阈值ω对实验结果的影响

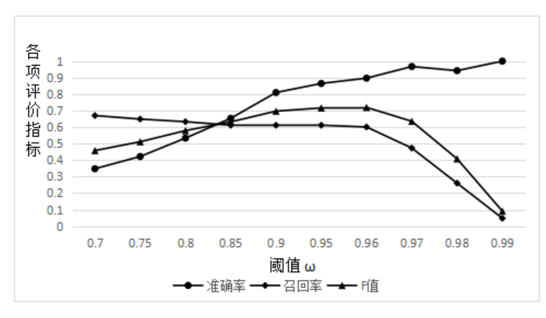
由上一次实验可知主题个数为8或9时各项指标最优，在以下实验中，主题个数K设定为9。实验结果如图6所示。



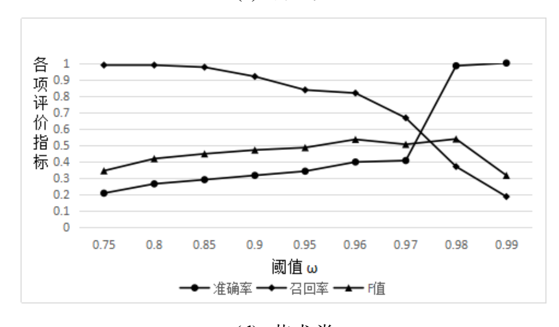
(a) 人物类



(b) 社会类



(c) 科学类



(d) 艺术类

图6 阈值ω对实验结果的影响

由图6可知，随着阈值的增加，这四类数据的准确率都在不断地增加，然而召回率却在不断地降低，这是由于在不断增加阈值的过程中，经过本算法之后正确对齐的实体数N_r不断减小，而F值则是先随着阈值的增加而增长，其大约在ω=0.96时取得最大值，由此可以得出当阈值取0.96时，该算法对齐效果最优。

3.4 与其他算法进行比较

为了证明所提算法的确实有效，将利用同样的文本数据信息与其他的算法进行实验效果比对。分别是在本文所提框架中主题建模部分改为TF-IDF、利用BP算法估计LDA模型隐藏参数和利用Gibbs算法推断LDA模型隐藏参数这三种算法进行对比实验。实验结果如表2所示。

表 2　与其他算法比较结果

算法	人物类			社会类			艺术类			科学类		
	P	R	F	P	R	F	P	R	F	P	R	F
本算法	0.620	0.867	0.723	0.641	0.964	0.770	0.399	0.818	0.536	0.865	0.612	0.767
TF-IDF	0.615	0.887	0.726	0.752	0.921	0.828	0.321	0.931	0.477	0.883	0.765	0.820
LDA+BP	0.435	0.675	0.529	0.623	0.881	0.730	0.353	0.803	0.490	0.752	0.612	0.675
LDA+Gibbs	0.615	0.830	0.707	0.633	0.962	0.764	0.400	0.753	0.522	0.842	0.771	0.805

由表 2 可以看出，对于相同的文本数据，并且实验参数设置相同的情况下，各算法实体对齐的效果不同，并有较大差异。首先，就本算法来说，其效果虽然和预想的结果差异较大，但是算法的准确率确实是高于 LDA+BP 的，由此可以证明本算法对于 BP 算法的改进确实是有效的。其次，TF-IDF 的准确率较本算法来说略低，大概率是因为 TF-IDF 仅仅是考虑了词项的词频信息而没有考虑文档的潜在语义。再次，由实验结果来看，LDA+Gibbs[23]的各项指标均与本算法的结果大致相同，这为本算法的再次优化提供了新的研究方向。最后，由实验数据可以看出，本算法的各项性能指标相对于原始算法有较大的提高，本算法对于解决百科知识库实体对齐的问题有较好的效果。

4　结束语

近年来，互联网规模的增长导致网络上知识信息大量的集中，知识库作为知识信息的载体在人们的学习中起到了重要的作用。然而单一的知识库的知识覆盖率较低，就需要通过知识融合的方式将各类不同的知识库进行整合，本文所提出的基于 LDA 的百科知识库实体对齐算法能够有效的解决知识库实体对齐问题，可以将其实际应用于百科知识库实体对齐工作中。

在之后的工作中，将考虑使用更有效的对 LDA 模型进行参数估计的方法，例如 Gibbs 抽样，以及发掘更多的主题模型来提高文本相似度，使知识库实体对齐的效果更加完善。

参考文献：

[1] Bollacker K, Cook R, Tufts P. Freebase: a shared database of structured general human knowledge [C]// Proc of AAAI Conference on Artificial Intelligence. British Columbia Canada: IEEE Press. 2007: 1962-1963.

[2] Lehmann J. DBpedia: A large-scale, multilingual knowledge base extracted from wikipedia [J]. Semantic Web, 2015, 6 (2): 167-195.

[3] Suchanek F M, Kasneci G, Weikum G. Yago: a large ontology from Wikipedia and WordNet [J]. Web Semantics Science Services & Agents on the World Wide Web, 2008, 6 (3): 203-217.

[4] Philpot A, Hovy E, Patrick P. The omega ontology [C]// Proc of Ontolex Workshop at LJCNLP. USA: Prep Press. 2005: 59-66.

[5] Li Mingyang, Shi Yao, Wang Zhigang, *et al.* Building a Large-Scale Cross-Lingual Knowledge Base from Heterogeneous Online Wikis [C]// Proc of CCF Conference on Natural Language Processing and Chinese Computing. New York: Springer-Verlag. 2015: 413-420.

[6] 徐增林，盛泳潘，贺丽荣，等. 知识图谱技术综述 [J]. 电子科技大学学报, 2016, 45 (4): 589-606. (Xu Zenglin, Sheng Yongpan, He Lirong, *et al.* Review on Knowledge Graph Techinques [J]. Journal od University of Electronic Science and Technology of China, 2016, 45 (4): 589-606.)

[7] 刘康，张元哲，纪国良，等. 基于表示学习的知识库问答研究进展与展望 [J]. 自动化学报, 2016, 42 (6): 807-818. (Liu Kang, Zhang Yuanzhe, Ji Guoliang, *et al.* Representation Learning for Question Answering over Knowledge Base: An Overview [J]. ACTA Automatica Sinica, 2016, 42 (6): 807-818.)

[8] 王雪鹏，刘康，何世柱，等. 基于网络语义标签的多源知识库实体对齐算法 [J]. 计算机学报, 2017, 40 (3): 701-711. (Wang Xuepeng, Liu Kang, He Shizhu, *et al.* Mult-Source Knowledge Base Entity Alignment by Leveraging Semantic Tags [J]. Chinese Journal of Computers, 2017, 40 (3): 701-711.)

[9] 高艳红，李爱萍，段利国. 面向实体链接的多特征图模型实体消歧方法 [J]. 计算机应用研究, 2017, 34 (10): 2909-2914. (Gao Yanhong, Li Aiping, Duan Liguo. Entity disambiguation method based on multi-feature fusion graph model for entity linking [J]. Application Research of Computers, 2017, 34 (10): 2909-2914)

[10] Liu Shulin, Liu Kang, He Shizhu, *et al.* A probabilistic soft logic based approach to exploiting latent and global information in event classification [C]// Proc of the 30th AAAI Conference on Artificial Intelligence. Phoenix USA: AAAI Press. 2016: 2993-2999.

[11] Stoilos G, Venetis T, Stamou G. A fuzzy extension to the OWL 2 RL ontology language [J]. Computer Journal, 2015, 58 (11): 2956-2971.

[12] 张晓辉，蒋海华，邸瑞华. 基于属性权重的链接数据共指关系构建 [J]. 计算机科学, 2013, 40 (2): 40-43. (Zhang Xiaohui, Jiang Haihua, Di Ruihua. Property Weight Based Co-reference Resolution for Linked Data [J]. Computer Science, 2013, 40 (2): 40-43.)

[13] 黄峻福，李天瑞，贾真，等. 中文异构百科知识库实体对齐 [J]. 计算机应用, 2016, 36 (7): 1881-1886. (Huang Junfu, Li Tianrui, Jia Zhen, *et al.* Entity alignment of Chinese heterogeneous encyclopedia knowledge base

收稿日期：2018-05-24；**修回日期**：2018-07-09　**基金项目**：河北省自然科学基金资助项目（2015201142）

作者简介：刘振鹏（1966-），男，河北安国人，教授，博士，主要研究方向为大数据、网络信息安全、自然语言处理；贺梦洁（1992-），女，硕士研究生，主要研究方向为大数据、自然语言处理；张彬（1980-），男（通信作者），高级工程师，硕士研究生，主要研究方向为网络安全（zb@hbu.edu.cn）；董静（1992-），女，硕士研究生，主要研究方向为大数据、自然语言处理；徐建民（1966-），男，教授，博导，主要研究方向为信息检索、不确定信息处理.

[J]. Journal of Computer Applications, 2016, 36 (7): 1881-1886.)

[14] 杨秀璋. 实体和属性对齐方法的研究与实现 [D]. 北京: 北京理工大学, 2016. (Yang Xiuzhang. Research and Implementation on Entity Alignment and Attribute Alignment [D]. Beijing: Beijing Institute of Technology, 2016.)

[15] 张伟莉, 黄廷磊, 梁霄. 基于半监督协同训练的百科知识库实体对齐 [J]. 计算机与现代化, 2017 (12): 88-93. (Zhang Weili, Huang Tinglei, Liang Xiao. Instance alignment algorithm between encyclopedia based on semi-supervised co-training [J]. Computer and Modernization, 2017 (12): 88-93)

[16] 万静, 李琳, 严欢春, 等. 基于 VS-Adaboost 的实体对齐方法 [J]. 北京化工大学学报: 自然科学版, 2018, 45 (1): 72-77. (Wan Jing, Li Lin, Yan Huanchun, *et al.* An entity alignment approach based on the VS-Adaboost aligorithm [J]. Journal of Beijing University of Chemical Technology: Natural Science Edition, 2018, 45 (1): 72-77.)

[17] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. J Machine Learning Research Archive, 2003, 3: 993-1022.

[18] Andersen S K. Probabilistic reasoning in intelligent systems: networks of plausible inference [J]. Artificial Intelligence, 1991, 48 (1): 117-124.

[19] Zeng Jia, Cheung William K, Liu Jiming. Learning topic models by belief propagation. [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2013, 35 (5): 1121-1134.

[20] 常东亚, 严建峰, 杨璐. 基于中心词的上下文主题模型 [J]. 计算机应用研究, 2018, 35 (4) . (Chang Dongya, Yan Jianfeng, Yang Lu. Centroid-word based context topic model [J]. Application Research of Computers, 2018, 35 (4): 1005-1009.)

[21] 庄严, 李国良, 冯建华. 知识库实体对齐技术综述 [J]. 计算机研究与发展, 2016, 53 (1): 165-192. (Zhuang Yan, Li Guoliang, Feng Jianhua. A survey on entity alignment of knowledge base [J]. Journal of Computer Research and Development, 2016, 53 (1): 165-192.)

[22] Wallach H M, Mimno D M, Mccallum A. Rethinking LDA: why priors matter [J]. Advances in Neural Information Processing Systems, 2009, 23: 1973-1981.

[23] 张健伟. 主题模型 LDA 推理算法对比与改进研究 [D]. 苏州: 苏州大学, 2017. (Zhang Jianwei. Cpmparison and Improvement Studies of Topic Model LDA Inference Algorithms [D]. Soochow: Soochow University, 2017.)